

Probabilistic & Unsupervised Learning

Week 3: The EM algorithm

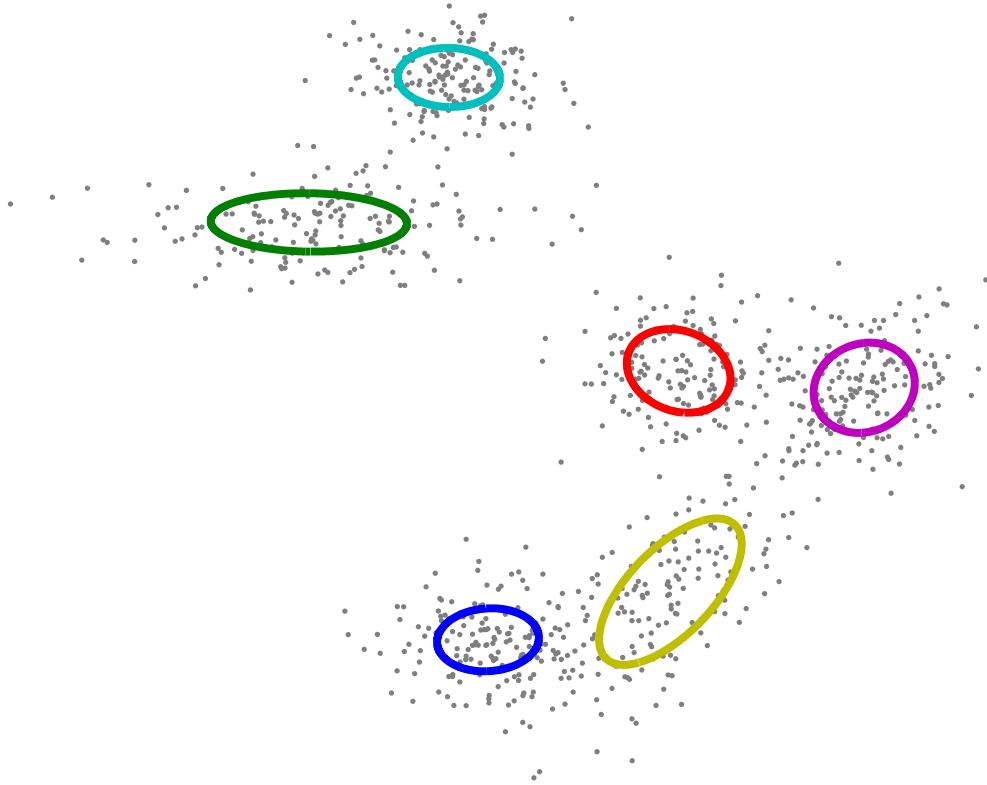
Maneesh Sahani

`maneesh@gatsby.ucl.ac.uk`

**Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London**

Term 1, Autumn 2010

Mixtures of Gaussians



Data: $\mathcal{X} = \{\mathbf{x}_1 \dots \mathbf{x}_N\}$

Latent process:

$$s_i \stackrel{\text{iid}}{\sim} \text{Disc}[\boldsymbol{\pi}]$$

Component distributions:

$$\mathbf{x}_i \mid (s_i = m) \sim \mathcal{P}_m[\theta_m] = \mathcal{N}(\boldsymbol{\mu}_m, \Sigma_m)$$

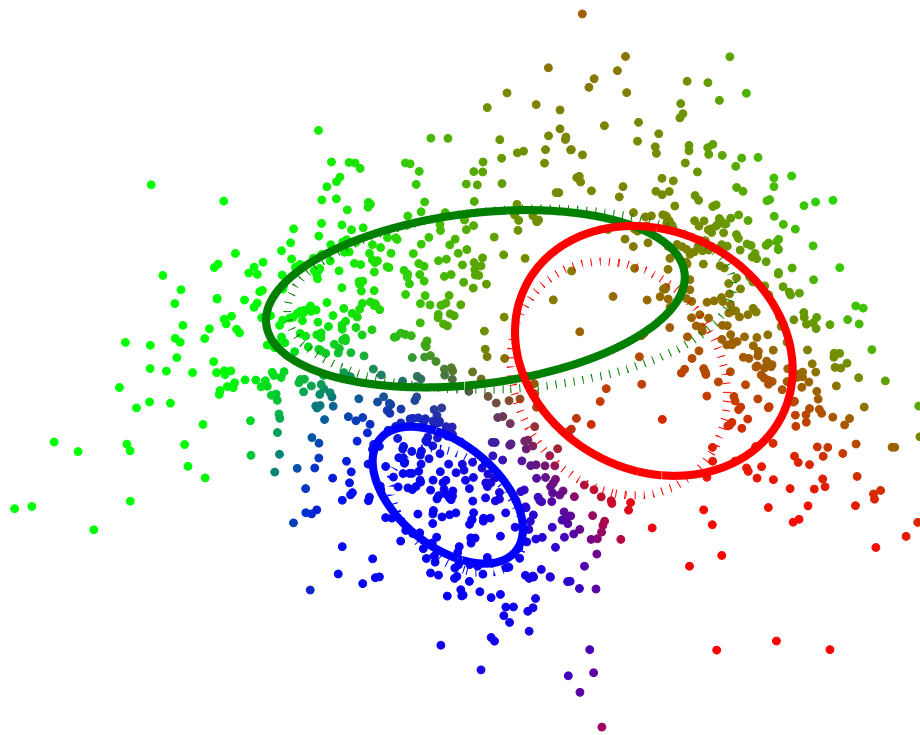
Marginal distribution:

$$P(\mathbf{x}_i) = \sum_{m=1}^k \pi_m P_m(\mathbf{x}; \theta_m)$$

Log-likelihood:

$$\log p(\mathcal{X} \mid \{\boldsymbol{\mu}_m\}, \{\Sigma_m\}, \boldsymbol{\pi}) = \sum_{i=1}^n \log \sum_{m=1}^k \pi_m |2\pi \Sigma_m|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_m)^\top \Sigma_m^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_m) \right]$$

EM for MoGs



- Evaluate responsibilities

$$r_{im} = \frac{P_m(\mathbf{x})\pi_m}{\sum_{m'} P_{m'}(\mathbf{x})\pi_{m'}}$$

- Update parameters

$$\boldsymbol{\mu}_m \leftarrow \frac{\sum_i r_{im} \mathbf{x}_i}{\sum_i r_{im}}$$

$$\Sigma_m \leftarrow \frac{\sum_i r_{im} (\mathbf{x}_i - \boldsymbol{\mu}_m)(\mathbf{x}_i - \boldsymbol{\mu}_m)^\top}{\sum_i r_{im}}$$

$$\pi_m \leftarrow \frac{\sum_i r_{im}}{N}$$

The Expectation Maximisation (EM) algorithm

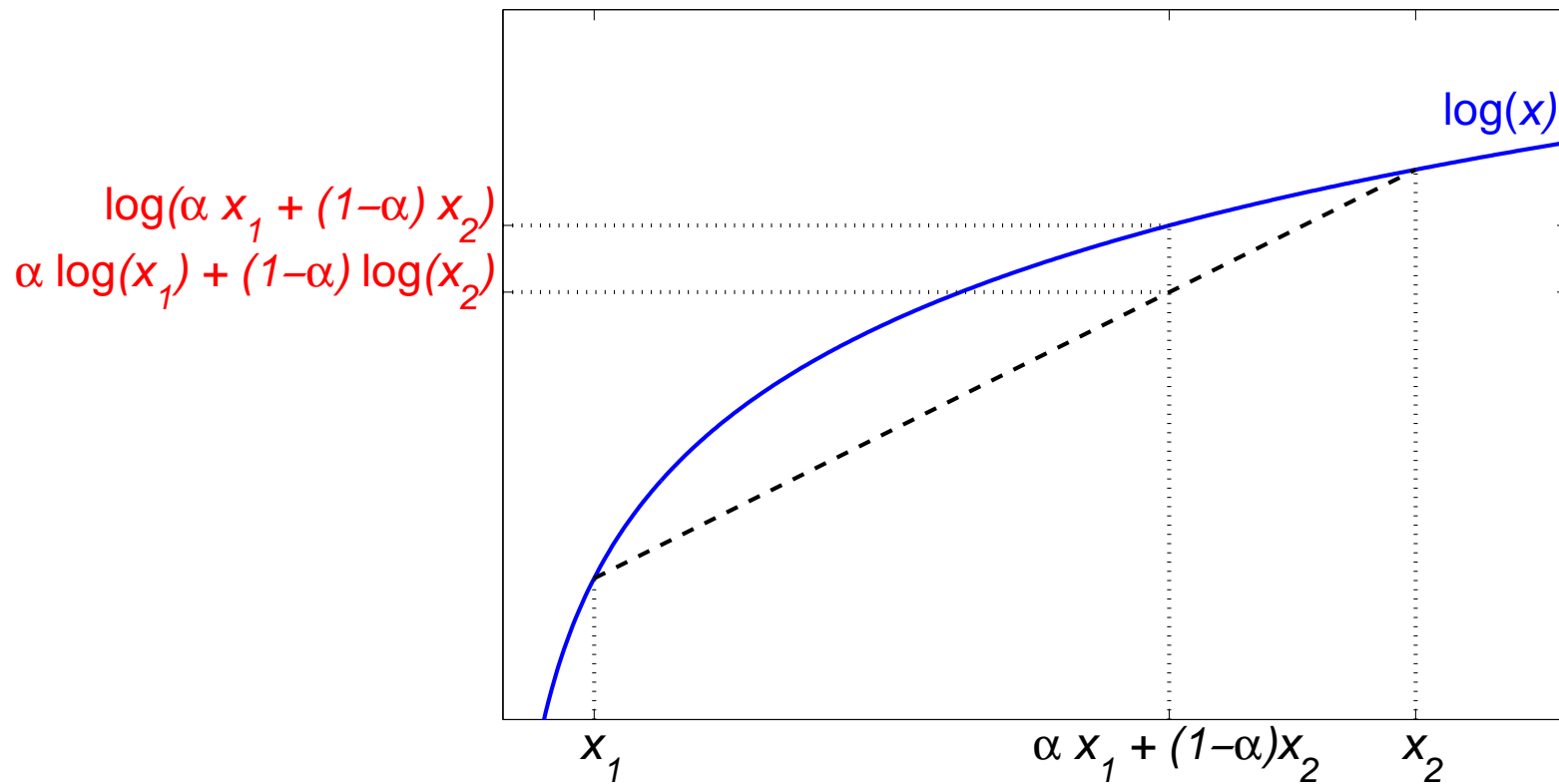
The EM algorithm finds a (local) maximum of a latent variable model likelihood. It starts from arbitrary values of the parameters, and iterates two steps:

E step: Fill in values of latent variables according to posterior given data.

M step: Maximise likelihood as if latent variables were not hidden.

- Useful in models where learning would be easy if hidden variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
- No learning rate.
- Framework lends itself to principled approximations.

Jensen's Inequality



For $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and any $\{x_i > 0\}$

$$\log \left(\sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i)$$

Equality if and only if $\alpha_i = 1$ for some i (and therefore all others are 0).

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ .

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ :

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} d\mathcal{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\begin{aligned} \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} d\mathcal{Y} &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y} + \mathbf{H}[q], \end{aligned}$$

where $\mathbf{H}[q]$ is the entropy of $q(\mathcal{Y})$.

So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + \mathbf{H}[q]$$

The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X} | \theta) \rangle_{q(\mathcal{Y})} + \mathbf{H}[q],$$

EM alternates between:

E step: optimize $\mathcal{F}(q, \theta)$ wrt distribution over hidden variables holding parameters fixed:

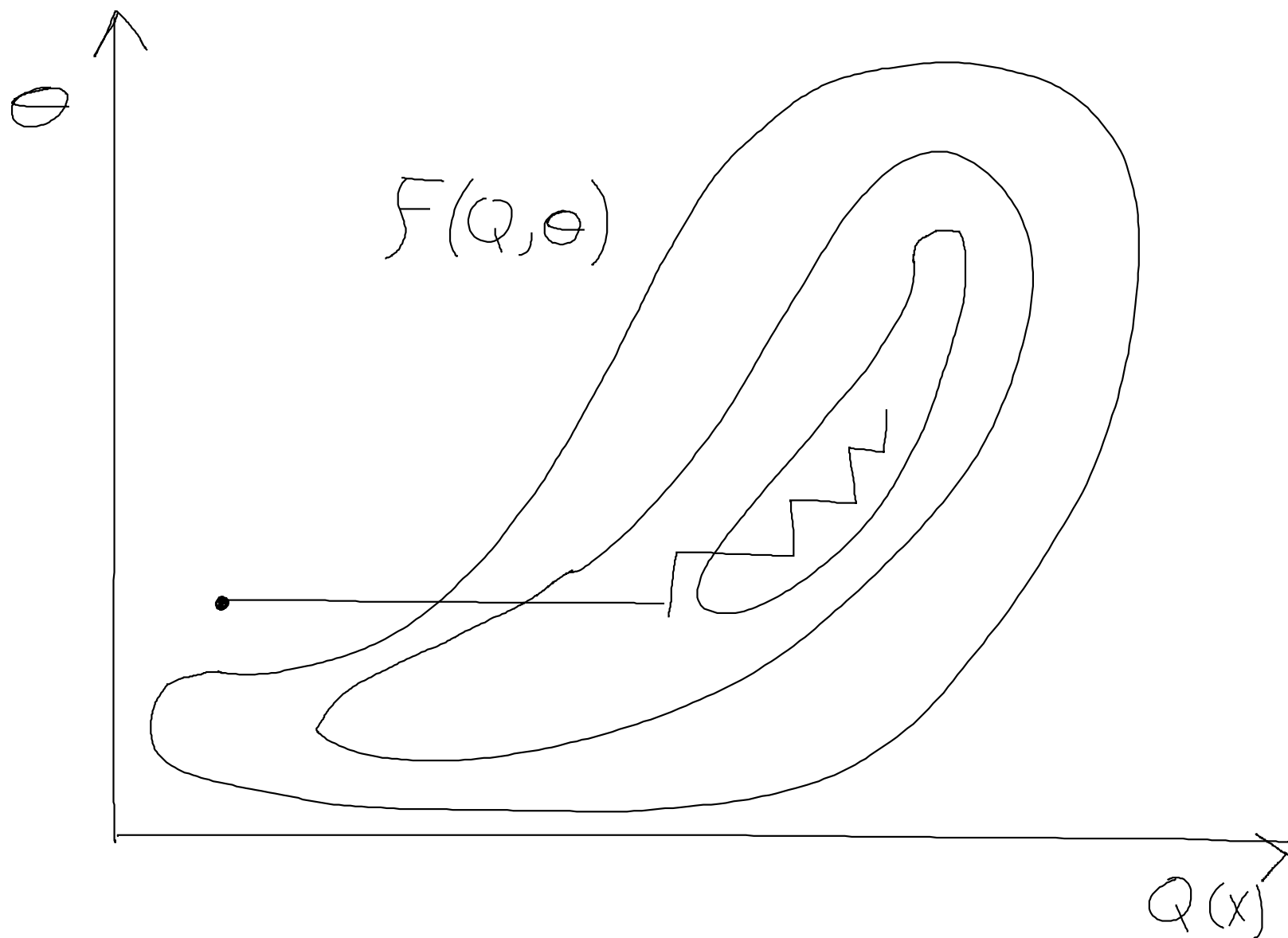
$$q^{(k)}(\mathcal{Y}) := \operatorname{argmax}_{q(\mathcal{Y})} \mathcal{F}(q(\mathcal{Y}), \theta^{(k-1)}).$$

M step: maximize $\mathcal{F}(q, \theta)$ wrt parameters holding hidden distribution fixed:

$$\theta^{(k)} := \operatorname{argmax}_{\theta} \mathcal{F}(q^{(k)}(\mathcal{Y}), \theta) = \operatorname{argmax}_{\theta} \langle \log P(\mathcal{Y}, \mathcal{X} | \theta) \rangle_{q^{(k)}(\mathcal{Y})}$$

The second equality comes from the fact that the entropy of $q(\mathcal{Y})$ does not depend directly on θ .

EM as Coordinate Ascent in \mathcal{F}



The E Step

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q, \theta) &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X} | \theta)}{q(\mathcal{Y})} d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y} | \mathcal{X}, \theta) P(\mathcal{X} | \theta)}{q(\mathcal{Y})} d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log P(\mathcal{X} | \theta) d\mathcal{Y} + \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y} | \mathcal{X}, \theta)}{q(\mathcal{Y})} d\mathcal{Y} \\ &= \ell(\theta) - \mathbf{KL}[q(\mathcal{Y}) \| P(\mathcal{Y} | \mathcal{X}, \theta)]\end{aligned}$$

The second term is the Kullback-Leibler divergence.

This means that, for fixed θ , \mathcal{F} is bounded above by ℓ , and achieves that bound when $\mathbf{KL}[q(\mathcal{Y}) \| P(\mathcal{Y} | \mathcal{X}, \theta)] = 0$.

But $\mathbf{KL}[q \| p]$ is zero if and only if $q = p$. So, the E step simply sets

$$q^{(k)}(\mathcal{Y}) = P(\mathcal{Y} | \mathcal{X}, \theta^{(k-1)})$$

and, after an E step, the free energy equals the likelihood.

The $\mathbf{KL}[q(x)||p(x)]$ is non-negative and zero iff $\forall x : p(x) = q(x)$

First let's consider discrete distributions; the Kullback-Liebler divergence is:

$$\mathbf{KL}[q||p] = \sum_i q_i \log \frac{q_i}{p_i}.$$

To find the distribution q which minimizes $\mathbf{KL}[q||p]$ we add a **Lagrange multiplier** to enforce the normalization constraint:

$$E \stackrel{\text{def}}{=} \mathbf{KL}[q||p] + \lambda(1 - \sum_i q_i) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda(1 - \sum_i q_i)$$

We then take partial derivatives and set to zero:

$$\left. \begin{aligned} \frac{\partial E}{\partial q_i} &= \log q_i - \log p_i + 1 - \lambda = 0 \Rightarrow q_i = p_i \exp(\lambda - 1) \\ \frac{\partial E}{\partial \lambda} &= 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1 \end{aligned} \right\} \Rightarrow q_i = p_i.$$

The $\mathbf{KL}[q(x)||p(x)]$ is non-negative and zero iff $\forall x : p(x) = q(x)$

Check that the curvature (Hessian) is positive (definite), corresponding to a minimum:

$$\frac{\partial^2 E}{\partial q_i \partial q_i} = \frac{1}{q_i} > 0, \quad \frac{\partial^2 E}{\partial q_i \partial q_j} = 0,$$

showing that $q_i = p_i$ is a genuine minimum.

At the minimum is it easily verified that $\mathbf{KL}[p||p] = 0$.

A similar proof holds for $\mathbf{KL}[\cdot||\cdot]$ between continuous densities, the derivatives being substituted by functional derivatives.

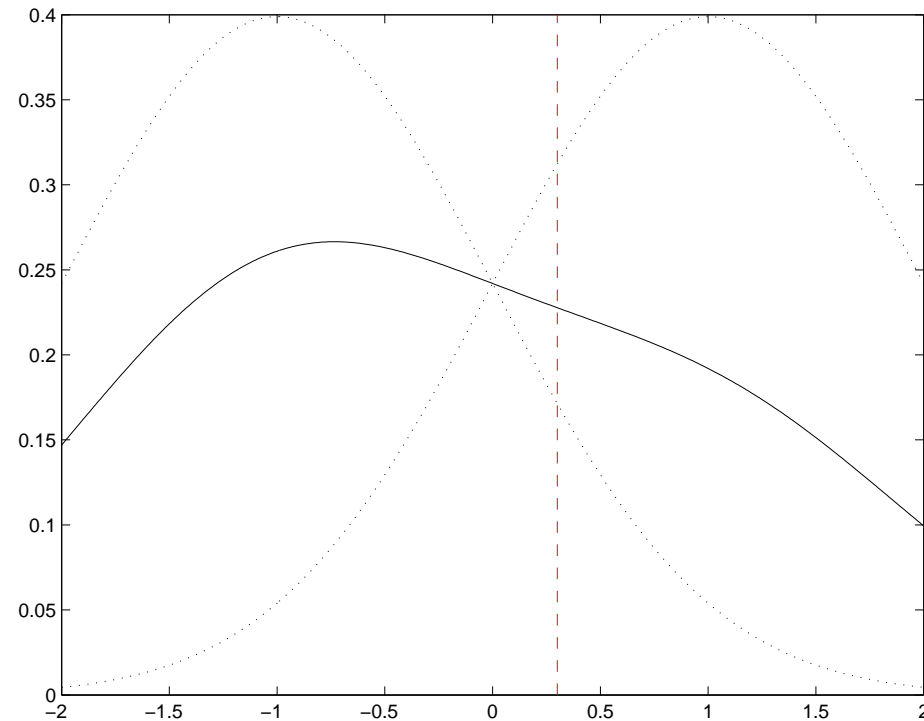
Coordinate Ascent in \mathcal{F} (Demo)

One parameter mixture:

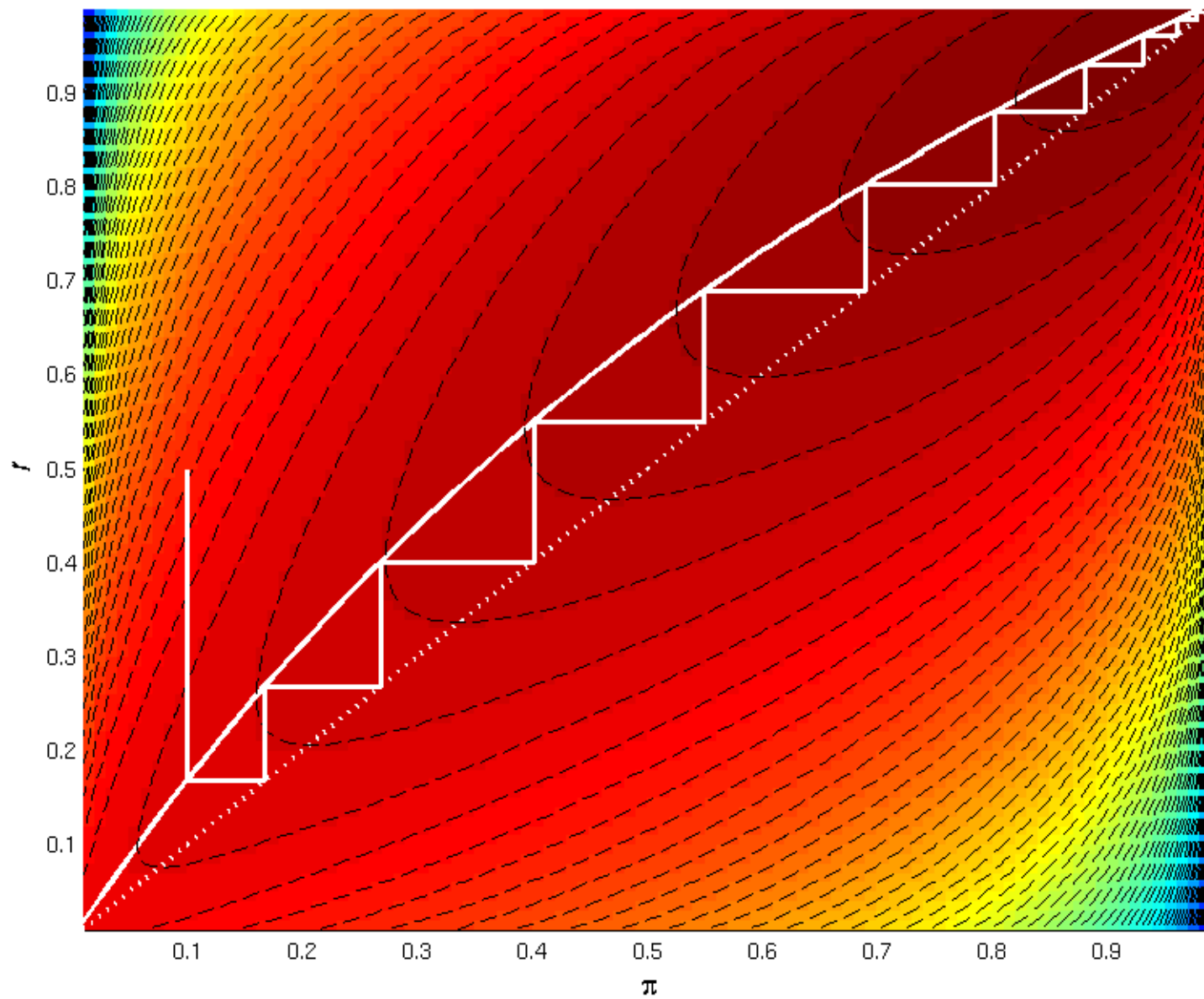
$$s \sim \text{Bernoulli}[\pi]$$
$$x|s=0 \sim \mathcal{N}[-1, 1] \quad x|s=1 \sim \mathcal{N}[1, 1]$$

and one data point $x_1 = .3$.

$q(s)$ is a distribution on a single binary latent, and so is represented by $r_1 \in [0, 1]$.



Coordinate Ascent in \mathcal{F} (Demo)



EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell(\theta^{(k-1)}) \underset{\text{E step}}{=} \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \underset{\text{M step}}{\leq} \mathcal{F}(q^{(k)}, \theta^{(k)}) \underset{\text{Jensen}}{\leq} \ell(\theta^{(k)}),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ .
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that $\theta^{(k)} \neq \theta^{(k-1)}$ iff \mathcal{F} increases, then the overall EM iteration will step to a new value of θ iff the likelihood increases.

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^* . Then:

$$\left. \frac{\partial}{\partial \theta} \langle \log P(\mathcal{Y}, \mathcal{X} | \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} \right|_{\theta^*} = 0$$

Now,

$$\begin{aligned} \ell(\theta) &= \log P(\mathcal{X} | \theta) = \langle \log P(\mathcal{X} | \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} \\ &= \left\langle \log \frac{P(\mathcal{Y}, \mathcal{X} | \theta)}{P(\mathcal{Y} | \mathcal{X}, \theta)} \right\rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} \\ &= \langle \log P(\mathcal{Y}, \mathcal{X} | \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} - \langle \log P(\mathcal{Y} | \mathcal{X}, \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} \end{aligned}$$

so,

$$\frac{d}{d\theta} \ell(\theta) = \frac{d}{d\theta} \langle \log P(\mathcal{Y}, \mathcal{X} | \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} - \frac{d}{d\theta} \langle \log P(\mathcal{Y} | \mathcal{X}, \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)}$$

The second term is 0 at θ^* if the derivative exists (minimum of $\mathbf{KL}[\cdot || \cdot]$), and thus:

$$\left. \frac{d}{d\theta} \ell(\theta) \right|_{\theta^*} = \left. \frac{d}{d\theta} \langle \log P(\mathcal{Y}, \mathcal{X} | \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} \right|_{\theta^*} = 0$$

So, EM converges to a stationary point of $\ell(\theta)$.

Maxima in \mathcal{F} correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of \mathcal{F} (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)} - \frac{d^2}{d\theta^2}\langle \log P(\mathcal{Y}|\mathcal{X}, \theta) \rangle_{P(\mathcal{Y}|\mathcal{X}, \theta^*)}$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum). Thus the curvature of the likelihood is negative and

θ^* is a maximum of ℓ .

[... as long as the derivatives exist. They sometimes don't (zero-noise ICA)].

Partial M steps and Partial E steps

Partial M steps: The proof holds even if we just *increase* \mathcal{F} wrt θ rather than maximize. (Dempster, Laird and Rubin (1977) call this the generalized EM, or GEM, algorithm).

Partial E steps: We can also just *increase* \mathcal{F} wrt to some of the q s.

For example, sparse or online versions of the EM algorithm would compute the posterior for a subset of the data points or as the data arrives, respectively. You can also update the posterior over a subset of the hidden variables, while holding others fixed...

The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point x is:

$$p(x|\theta) = \sum_{m=1}^k p(s = m|\theta)p(x|s = m, \theta) \propto \sum_{m=1}^k \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x - \mu_m)^2 \right\},$$

where θ is the collection of parameters: means μ_m , variances σ_m^2 and mixing proportions $\pi_m = p(s = m|\theta)$.

The hidden variable s_i indicates which component observation x_i belongs to.

The E-step computes the posterior for s_i given the current parameters:

$$q(s_i) = p(s_i|x_i, \theta) \propto p(x_i|s_i, \theta)p(s_i|\theta)$$
$$r_{im} \stackrel{\text{def}}{=} q(s_i = m) \propto \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right\} \quad (\text{responsibilities})$$

with the normalization such that $\sum_m r_{im} = 1$.

The Gaussian mixture model (M-step)

In the M-step we optimize the sum (since s is discrete):

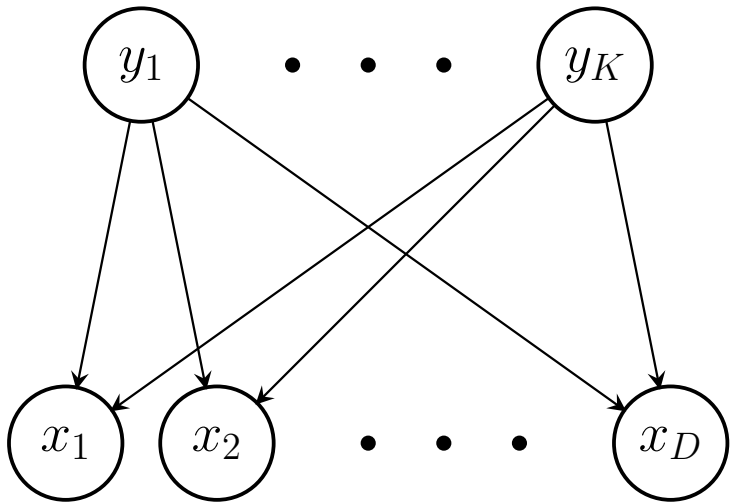
$$\begin{aligned} E &= \langle \log p(x, s | \theta) \rangle_{q(s)} = \sum q(s) \log[p(s | \theta) p(x | s, \theta)] \\ &= \sum_{i,m} r_{im} \left[\log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right]. \end{aligned}$$

Optimization is done by setting the partial derivatives of E to zero:

$$\begin{aligned} \frac{\partial E}{\partial \mu_m} &= \sum_i r_{im} \frac{(x_i - \mu_m)}{2\sigma_m^2} = 0 \Rightarrow \mu_m = \frac{\sum_i r_{im} x_i}{\sum_i r_{im}}, \\ \frac{\partial E}{\partial \sigma_m} &= \sum_i r_{im} \left[-\frac{1}{\sigma_m} + \frac{(x_i - \mu_m)^2}{\sigma_m^3} \right] = 0 \Rightarrow \sigma_m^2 = \frac{\sum_i r_{im} (x_i - \mu_m)^2}{\sum_i r_{im}}, \\ \frac{\partial E}{\partial \pi_m} &= \sum_i r_{im} \frac{1}{\pi_m}, \quad \frac{\partial E}{\partial \pi_m} + \lambda = 0 \Rightarrow \pi_m = \frac{1}{n} \sum_i r_{im}, \end{aligned}$$

where λ is a Lagrange multiplier ensuring that the mixing proportions sum to unity.

Factor Analysis



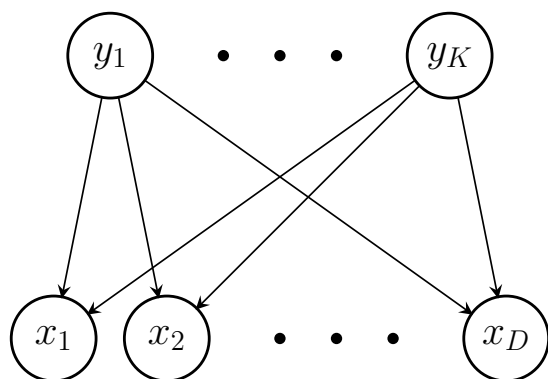
Linear generative model:
$$x_d = \sum_{k=1}^K \Lambda_{dk} y_k + \epsilon_d$$

- y_k are independent $\mathcal{N}(0, 1)$ Gaussian **factors**
- ϵ_d are independent $\mathcal{N}(0, \Psi_{dd})$ Gaussian **noise**
- $K < D$

So, \mathbf{x} is Gaussian with: $p(\mathbf{x}) = \int p(\mathbf{y})p(\mathbf{x}|\mathbf{y})d\mathbf{y} = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$
where Λ is a $D \times K$ matrix, and Ψ is diagonal.

Dimensionality Reduction: Finds a low-dimensional projection of high dimensional data that captures the **correlation structure** of the data.

EM for Factor Analysis



The model for \mathbf{x} :

$$p(\mathbf{x}|\theta) = \int p(\mathbf{y}|\theta)p(\mathbf{x}|\mathbf{y}, \theta)d\mathbf{y} = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$$

Model parameters: $\theta = \{\Lambda, \Psi\}$.

E step: For each data point \mathbf{x}_n , compute the posterior distribution of hidden factors given the observed data: $q_n(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}_n, \theta_t)$.

M step: Find the θ_{t+1} that maximises $\mathcal{F}(q, \theta)$:

$$\begin{aligned}\mathcal{F}(q, \theta) &= \sum_n \int q_n(\mathbf{y}) [\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y}, \theta) - \log q_n(\mathbf{y})] d\mathbf{y} \\ &= \sum_n \int q_n(\mathbf{y}) [\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y}, \theta)] d\mathbf{y} + \mathbf{c}.\end{aligned}$$

The E step for Factor Analysis

E step: For each data point \mathbf{x}_n , compute the posterior distribution of hidden factors given the observed data: $q_n(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}_n, \theta) = p(\mathbf{y}, \mathbf{x}_n|\theta)/p(\mathbf{x}_n|\theta)$

Tactic: write $p(\mathbf{y}, \mathbf{x}_n|\theta)$, consider \mathbf{x}_n to be fixed. What is this as a function of \mathbf{y} ?

$$\begin{aligned} p(\mathbf{y}, \mathbf{x}_n) &= p(\mathbf{y})p(\mathbf{x}_n|\mathbf{y}) \\ &= (2\pi)^{-\frac{K}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}^\top \mathbf{y}\right\} |2\pi\Psi|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_n - \Lambda\mathbf{y})^\top \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y})\right\} \\ &= \mathbf{c} \times \exp\left\{-\frac{1}{2}[\mathbf{y}^\top \mathbf{y} + (\mathbf{x}_n - \Lambda\mathbf{y})^\top \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y})]\right\} \\ &= \mathbf{c}' \times \exp\left\{-\frac{1}{2}[\mathbf{y}^\top (I + \Lambda^\top \Psi^{-1} \Lambda) \mathbf{y} - 2\mathbf{y}^\top \Lambda^\top \Psi^{-1} \mathbf{x}_n]\right\} \\ &= \mathbf{c}'' \times \exp\left\{-\frac{1}{2}[\mathbf{y}^\top \Sigma^{-1} \mathbf{y} - 2\mathbf{y}^\top \Sigma^{-1} \mu + \mu^\top \Sigma^{-1} \mu]\right\} \end{aligned}$$

So $\Sigma = (I + \Lambda^\top \Psi^{-1} \Lambda)^{-1} = I - \beta \Lambda$ and $\mu = \Sigma \Lambda^\top \Psi^{-1} \mathbf{x}_n = \beta \mathbf{x}_n$. Where $\beta = \Sigma \Lambda^\top \Psi^{-1}$.
Note that μ is a linear function of \mathbf{x}_n and Σ does not depend on \mathbf{x}_n .

The M step for Factor Analysis

M step: Find θ_{t+1} maximising $\mathcal{F} = \sum_n \int q_n(\mathbf{y}) [\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y}, \theta)] d\mathbf{y} + \mathbf{c}$

$$\begin{aligned}\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y}, \theta) &= \mathbf{c} - \frac{1}{2}\mathbf{y}^\top \mathbf{y} - \frac{1}{2} \log |\Psi| - \frac{1}{2}(\mathbf{x}_n - \Lambda \mathbf{y})^\top \Psi^{-1}(\mathbf{x}_n - \Lambda \mathbf{y}) \\ &= \mathbf{c}' - \frac{1}{2} \log |\Psi| - \frac{1}{2}[\mathbf{x}_n^\top \Psi^{-1} \mathbf{x}_n - 2\mathbf{x}_n^\top \Psi^{-1} \Lambda \mathbf{y} + \mathbf{y}^\top \Lambda^\top \Psi^{-1} \Lambda \mathbf{y}] \\ &= \mathbf{c}' - \frac{1}{2} \log |\Psi| - \frac{1}{2}[\mathbf{x}_n^\top \Psi^{-1} \mathbf{x}_n - 2\mathbf{x}_n^\top \Psi^{-1} \Lambda \mathbf{y} + \text{Tr} [\Lambda^\top \Psi^{-1} \Lambda \mathbf{y} \mathbf{y}^\top]]\end{aligned}$$

Taking expectations over $q_n(\mathbf{y}) \dots$

$$= \mathbf{c}' - \frac{1}{2} \log |\Psi| - \frac{1}{2}[\mathbf{x}_n^\top \Psi^{-1} \mathbf{x}_n - 2\mathbf{x}_n^\top \Psi^{-1} \Lambda \mu_n + \text{Tr} [\Lambda^\top \Psi^{-1} \Lambda (\mu_n \mu_n^\top + \Sigma)]]$$

Note that we don't need to know everything about q , just the expectations of \mathbf{y} and $\mathbf{y} \mathbf{y}^\top$ under q (i.e. the expected sufficient statistics).

The M step for Factor Analysis (cont.)

$$\mathcal{F} = c - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n [\mathbf{x}_n^\top \Psi^{-1} \mathbf{x}_n - 2\mathbf{x}_n^\top \Psi^{-1} \Lambda \mu_n + \text{Tr} [\Lambda^\top \Psi^{-1} \Lambda (\mu_n \mu_n^\top + \Sigma)]]$$

Taking derivatives w.r.t. Λ and Ψ^{-1} , using $\frac{\partial \text{Tr}[AB]}{\partial B} = A^\top$ and $\frac{\partial \log |A|}{\partial A} = A^{-\top}$:

$$\frac{\partial \mathcal{F}}{\partial \Lambda} = \Psi^{-1} \sum_n \mathbf{x}_n \mu_n^\top - \Psi^{-1} \Lambda \left(N\Sigma + \sum_n \mu_n \mu_n^\top \right) = 0$$

$$\hat{\Lambda} = \left(\sum_n \mathbf{x}_n \mu_n^\top \right) \left(N\Sigma + \sum_n \mu_n \mu_n^\top \right)^{-1}$$

$$\frac{\partial \mathcal{F}}{\partial \Psi^{-1}} = \frac{N}{2} \Psi - \frac{1}{2} \sum_n [\mathbf{x}_n \mathbf{x}_n^\top - \Lambda \mu_n \mathbf{x}_n^\top - \mathbf{x}_n \mu_n^\top \Lambda^\top + \Lambda (\mu_n \mu_n^\top + \Sigma) \Lambda^\top]$$

$$\hat{\Psi} = \frac{1}{N} \sum_n [\mathbf{x}_n \mathbf{x}_n^\top - \Lambda \mu_n \mathbf{x}_n^\top - \mathbf{x}_n \mu_n^\top \Lambda^\top + \Lambda (\mu_n \mu_n^\top + \Sigma) \Lambda^\top]$$

$$\hat{\Psi} = \Lambda \Sigma \Lambda^\top + \frac{1}{N} \sum_n (\mathbf{x}_n - \Lambda \mu_n)(\mathbf{x}_n - \Lambda \mu_n)^\top \quad (\text{squared residuals})$$

Note: we should actually only take derivatives w.r.t. Ψ_{dd} since Ψ is diagonal.
When $\Sigma \rightarrow 0$ these become the equations for linear regression!

Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

$$p(\mathbf{x}|\theta) = \sum_k \pi_k \mathcal{N}(\mu_k, \Lambda_k \Lambda_k^\top + \Psi)$$

where π_k is the mixing proportion for FA k , μ_k is its centre, Λ_k is its “factor loading matrix”, and Ψ is a common sensor noise model. $\theta = \{\{\pi_k, \mu_k, \Lambda_k\}_{k=1\dots K}, \Psi\}$

We can think of this model as having *two* sets of hidden latent variables:

- A discrete indicator variable $s_n \in \{1, \dots, K\}$
- For each factor analyzer, a continuous factor vector $\mathbf{y}_{n,k} \in \mathcal{R}^{D_k}$

$$p(\mathbf{x}|\theta) = \sum_{s_n=1}^K p(s_n|\theta) \int p(\mathbf{y}|s_n, \theta) p(\mathbf{x}_n|\mathbf{y}, s_n, \theta) d\mathbf{y}$$

As before, an EM algorithm can be derived for this model:

E step: Infer joint distribution of latent variables, $p(\mathbf{y}_n, s_n|\mathbf{x}_n, \theta)$

M step: Maximize \mathcal{F} with respect to θ .

EM for exponential families

Defn: p is in the exponential family for $\mathbf{z} = (\mathbf{y}, \mathbf{x})$ if it can be written:

$$p(\mathbf{z}|\theta) = b(\mathbf{z}) \exp\{\theta^\top s(\mathbf{z})\} / \alpha(\theta)$$

where $\alpha(\theta) = \int b(\mathbf{z}) \exp\{\theta^\top s(\mathbf{z})\} d\mathbf{z}$

E step: $q(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \theta)$

M step: $\theta^{(k)} := \operatorname{argmax}_{\theta} \mathcal{F}(q, \theta)$

$$\begin{aligned} \mathcal{F}(q, \theta) &= \int q(\mathbf{y}) \log p(\mathbf{y}, \mathbf{x}|\theta) d\mathbf{y} - \mathcal{H}(q) \\ &= \int q(\mathbf{y}) [\theta^\top s(\mathbf{z}) - \log \alpha(\theta)] d\mathbf{y} + \text{const} \end{aligned}$$

It is easy to verify that: $\frac{\partial \log \alpha(\theta)}{\partial \theta} = E[s(\mathbf{z})|\theta]$

Therefore, M step solves: $\frac{\partial \mathcal{F}}{\partial \theta} = E_{q(\mathbf{y})}[s(\mathbf{z})] - E[s(\mathbf{z})|\theta] = 0$

References

- A. P. Dempster, N. M. Laird and D. B. Rubin (1977).
Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1 (1977), pp. 1-38.
<http://www.jstor.org/stable/2984875>
- R. M. Neal and G. E. Hinton (1998).
A view of the EM algorithm that justifies incremental, sparse, and other variants.
In M. I. Jordan (editor) Learning in Graphical Models, pp. 355-368, Dordrecht: Kluwer Academic Publishers.
<http://www.cs.utoronto.ca/~radford/ftp/emk.pdf>
- Z. Ghahramani and G. E. Hinton (1996).
The EM Algorithm for Mixtures of Factor Analyzers.
University of Toronto Technical Report CRG-TR-96-1.
<http://learning.eng.cam.ac.uk/zoubin/papers/tr-96-1.pdf>

Proof of the Matrix Inversion Lemma

$$(A + XBX^\top)^{-1} = A^{-1} - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}$$

Need to prove:

$$(A^{-1} - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1})(A + XBX^\top) = I$$

Expand:

$$I + A^{-1}XBX^\top - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}XBX^\top$$

Regroup:

$$\begin{aligned} &= I + A^{-1}X(BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}X^\top - (B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}XBX^\top) \\ &= I + A^{-1}X(BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}B^{-1}BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}XBX^\top) \\ &= I + A^{-1}X(BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}(B^{-1} + X^\top A^{-1}X)BX^\top) \\ &= I + A^{-1}X(BX^\top - BX^\top) = I \end{aligned}$$